METHOD FOR GENERATING DATA RECORDS FROM A DATABASE, ESPECIALLY FROM THE WORLD WIDE WEB, CHARACTERISTIC SHORT DATA RECORDS, METHOD FOR DETERMINING DATA RECORDS FROM A DATABASE WHICH ARE RELEVANT FOR A PREDEFINED SEARCH QUERY AND SEARCHING SYSTEM FOR IMPLEMENTING SAID METHOD

Description

[0001] The present invention relates to a method for generating short data records that are characteristic of data records from a database, especially from the World Wide Web, for storage in a memory module as the basis for determining the data records that are relevant for a specifiable search query. The invention also relates to a method for determining data records that are relevant for a specifiable search query from a database, particularly from the World Wide Web, in which method such short data records are searched for their relevancy for the particular search query. Moreover, the invention relates to a searching system for determining data records that are relevant for a specifiable search query from a database, especially from the World Wide Web.

[0002] Complex databases or else the worldwide computer network (World Wide Web) contain an enormous volume of information that can be retrieved by a user in a more or less systematic manner for searching purposes. In order to be able to systematically render information useable from among this large volume of information, so-called search engines are employed, some of which now enjoy widespread use, especially when it comes to acquiring information from the World Wide Web. With such search engines, the user is normally offered a query window via an input/output module by means of which targeted searches or search terms can be entered. Subsequently, the search engine browses the information base of the database or the World Wide Web for suitable key words. The answer data records thus found are normally categorized by the appertaining search engine in terms of their relevancy for the specified search query and are then made available to the user in the form of a list of hits arranged in the order of their relevancy.

[0003] However, the increasing complexity of the databases employed and especially the ever-growing abundance of information content on the World Wide Web pose increasing problems for the structured and systematic acquisition of information. As a consequence, the search engines used for searching are constantly being improved in terms of the search algorithms they employ, whereby additional resources along the lines of pre-sorting or pre-filtering can be used to classify data records from the database. Particularly on the World Wide Web, data records are normally structured and organized in the form of so-called domains, such a domain typically being maintained by an operator and being capable, in turn, of encompassing numerous sub-data records, text documents and the like.

[0004] With an eye towards being able to make an appropriate pre-selection of the domains that are to be taken into consideration for a given search query, especially when information is to be acquired from the World Wide Web, despite the enormous number of data records and domains available there containing sub-data records or information carriers whose contents might be relatively large, a so-called ranking of the domains can be employed. In this process, each domain is associated with a characteristic value which, on the basis of accessible secondary information in the form of a relative relevancy, characterizes the importance of taking that particular domain into consideration for the search query. When this characteristic value is associated with a given domain, an information base in the form of a so-called static approach is normally employed in which a conclusion is drawn about the relative significance of this particular domain, for example, on the basis of the degree of networking of said domain with other domains. The number of so-called links or cross references from other domains to this particular domain can serve as a measure of such significance, whereby the assumption is made that a large number of cross references to a given domain is an indication that this domain is particularly important for many users when their search queries are being processed.

[0005] However, it has been found that, when such a static characteristic value is ascribed as an indicator of relevancy for a given domain, there is room for manipulations whereby, irrespective of the actual interest on the part of the user, financial considerations can give rise to the generation of a large number of actually unwarranted links or cross references that artificially create the impression of a relatively high relevancy or significance of a given domain. Therefore,

the use of such static relevancy associations in order to improve the search results of Internet searches is on the decline.

[0006] When an information search is carried out, the huge volumes of information available on the World Wide Web or Internet make it impossible to actually search all of the domains, including the sub-data records or text blocks, in real time for the presence of the search query or of individual elements of a given search query. Instead, so-called "crawlers" or browsing modules are employed in searching systems or search engines to acquire information from the Internet or World Wide Web, said crawlers carrying out a continuous search of the domains or data records from the World Wide Web or from a complex database for their text contents or other information deemed to be relevant. Within the scope of predefined system resources (for instance, processing time, storage capacity or computing capacity), the browsing module in question browses the currently selected domain or data record and, up to a limit specified by the allocated system resources and on the basis of the information found in that particular domain, compiles a short data record – for example, in the form of a text file that might have associated headings or other indicators – that is characteristic for the domain or data record.

[0007] This short data record is then stored in a memory module and kept ready for a subsequent search. The totality of the short data records that are generated from the data records or domains that are actually taken into consideration in this process and that are stored on the memory module is also designated as a so-called "index" of the search engine in question and serves as the information base for the subsequently performed searches. The short data records contained in the index are then normally generated continuously, whereby individual domains are cyclically selected so that the index is updated on an ongoing basis. During a subsequent search, in other words, when the data records that are relevant for a specified search query are being determined, the index formed by the totality of the stored short data records is searched for the presence of key words of a given search query or of individual elements of said search query, whereby the search results or hits thus obtained are used to ascertain the data records or domains that are associated with the short data records found as being relevant for the particular search query.

[0008] In view of the number of domains or data records available on the World Wide Web, it is not possible to consider all of the domains when the short data records are generated. The decision as to which domains are taken into consideration for generating the index is normally made on the basis of the above-mentioned relevancy criteria, in other words, especially on the basis of information regarding the recognized or presumed significance of a given domain for the user. Precisely because of the sheer abundance of information available, it can be very important to undertake an especially systematic pre-sorting of the information and particularly of the data records that have been recognized as being relevant for a search query within the scope of the subsequent evaluation of the search results, and consequently, already when the so-called index is being generated, it is desirable to have a particularly high level of quality and thoroughness in the evaluation of the information that has been considered.

[0009] Therefore, the invention is based on the objective of putting forward a method for generating short data records that are characteristic of data records of the above-mentioned type, by means of which method it is possible to generate a search index that is particularly well-suited for acquiring high-quality information from the database or from the World Wide Web. Furthermore, with the use of said method, a particularly suitable method for determining data records that are relevant for a specifiable search query from a database, especially from the World Wide Web, and a searching system to carry out this method are also to be put forward.

[0010] Regarding the method for generating short data records that are characteristic of the data records, this objective is achieved according to the invention in that the system resources provided for generating a short data record from a data record are selected taking into consideration empirical values determined in preceding search queries.

[0011] In this context, the invention is based on the consideration that, in order to generate an information base that is especially well-suited for acquiring particularly high-quality information based on the short data records that are characteristic of the data records, it is possible, on the one hand, to take into consideration information about the individual data records or domains that is available in the form of static characteristic values while, on the other hand, it is likewise necessary to consider information in the form of a dynamic element that is

also characteristic of the user interests. This is founded on the realization that the result of an acquisition of information from the database or from the World Wide Web can be regarded as being of a particularly high quality if it correctly reflects the user interests to the greatest possible extent. Therefore, measures should be taken so as to allow information that is characteristic of the user interests to be integrated into the further acquisition of information. An approach towards doing this is already the generation of the information base for processing search queries so that information about user interests should already be reflected in the index when the short data records that are characteristic of the data records or domains are generated. In order to permit this, with an eye towards the user interests and taking into account empirical values from preceding search queries, resources are already allocated at the time of allocation of the system resources that are used to generate a short data record from an associated data record and that play a decisive role in determining the completeness of the information made available in the short data record for the acquisition of information.

[0012] In a particularly simple and effective manner, the user interests can already be taken into consideration when the index is generated in that, advantageously during the allocation of the system resources, the frequency of recent search queries that are identical or similar to a given search query is taken into account as an empirical value. In another advantageous embodiment, the frequency of hits of the data records or domains pertaining to the search queries that have recently been specified particularly often by users is taken into consideration. Consequently, the empirical values advantageously encompass a characteristic value that is characteristic of the number of similar search queries within a specifiable time span.

[0013] In order to be able to take the user interests into consideration in a particularly targeted manner already during the generation of the index for the search engine, the resources of a browsing module or crawler provided for generating the short data records that are characteristic of the data records are advantageously selected as the system resources, also taking into consideration empirical values obtained from previous search queries.

[0014] In an especially advantageous embodiment, the user interests are taken into consideration to a particularly great extent during the allocation of the system resources in that,

when the empirical values are determined, special attention is paid to the possibly complex structure of the search queries entered by the users. This is based on the realization that a particularly accurate notion of the general interests of users can be obtained not only on the basis of the relative frequency of individual elements or terms employed in the search queries but also, as a supplement or in addition, by considering specific correlations between individual terms or elements of the search queries. Here, special attention is paid to the fact that individual elements or components of a search query are requested on the basis of the user interests that are currently widespread, preferably in combination with specific other individual elements or components of search queries. For instance, the current user interest might generally point in the direction that preferably free multimedia files should be downloaded from the Internet. Before the backdrop of such a constellation, it can be expected that search queries will more often contain a combination of the search terms "MP3", "free" and "download". Therefore, in the specific evaluation and consideration of past search queries, the combination of these three individual elements of a search query can be assessed as being a particularly significant indicator of heightened user interest. In order to make this possible, preferably correlations between individual elements of the search queries are considered when the empirical values are being determined.

[0015] For purposes of providing initial information that is relatively easy to obtain for the evaluation of search queries and their relevancy for the data records, along the lines of a first pre-filtering, the relative frequency of search queries and/or of individual elements of the search queries are advantageously taken into account when the empirical values are determined. This can be taken into consideration particularly easily directly during the generation of the index in that the data records that are recognized as being relevant for a specified search query or for a specified combination of individual elements of search queries advantageously receive an allocation of additional system resources – as a function of the relative frequency of the search query or of the combination of individual elements of search queries – for the generation of the associated short data record.

[0016] Advantageously, the short data records that are characteristic of the data records from the database and that are generated in the above-mentioned manner are used to determine data records from the database, especially from the World Wide Web, which are relevant for a

specifiable search query; this is done in that the short data records thus generated and stored in a memory module are searched for their relevancy for a given search query. The criterion for determining this relevancy can be, for example, the frequency with which a key word of the search query can be found in a particular short data record, whereby a differentiation can also be made on the basis of the location of the found passage, for instance, in a heading or in the running text.

[0017] Regarding the searching system for determining data records from database, especially from the World Wide Web, which are relevant for a specifiable search query, the above-mentioned objective is achieved in that short data records that are characteristic of the data records are stored in a memory module, whereby the system resources provided for the generation of a short data record from a data record are selected taking into consideration stored empirical values from preceding search queries.

[0018] These empirical values advantageously encompass a characteristic number that is characteristic of the number of similar search queries within a specifiable time span. In an additional or alternative advantageous embodiment, the resources of a browsing module provided for the generation of the short data records that are characteristic of the data records are selected as the system resources, taking into consideration stored empirical values from preceding search queries.

[0019] The advantages achieved with the invention lie primarily in the fact that, by taking into account empirical values from preceding search queries when the system resources are allocated during the generation of the index or of the short data records that are characteristic of the data records, the currently applicable user interests are largely taken into consideration already at a particularly early point in time, namely, during the preparation phase of a database or Internet search. Precisely as a result of considering user interests in addition to or instead of the database-specific characteristics employed so far, such as, for instance, the frequency of particular cross references, it is now possible to acquire information that is recognized as being of a particularly high quality by the user. A particularly specific indication of the interests of the user and thus an especially accurate allocation of the resources can be obtained by taking into

account correlations between individual elements of search queries, whereby precisely very frequently employed combinations of specific individual elements and the link back to the data records or domains found as results with such combined search queries means that hits can be generated that match the user interest to a high degree.

[0020] An embodiment of the invention will be explained in greater detail with reference to a drawing. In this drawing, the figure schematically shows a searching system for determining data records or domains from the World Wide Web that are relevant for a specifiable search query.

[0021] The searching system 1 according to the figure is connected to a plurality of domains 4 via the data lines of the Internet or of the World Wide Web indicated by the double-headed arrow 2; each domain 4, in turn, typically comprises numerous sub-data records, text blocks, multimedia information elements and the like.

[0022] Owing to the large amount of information available on the World Wide Web, the searching system 1 for processing a search query is not configured to browse the domain 4 or the information contained there for the presence of certain key words, but rather, it is configured to browse a so-called index 8 that is stored in a memory module 6. The index 8 encompasses numerous short data records 10, each of which is characteristic of a data record or domain 4 of the World Wide Web. Each short data record 10 contains a part of the information content recognized as being relevant in the associated domain 4, whereby especially the text information contained in the appertaining domain 4 is found in the short data record 10. In order to process a search query, the latter, as indicated by the arrow 12, is fed to an input/output module 14 of the searching system 1, from where the browsing of the short data records 10 is started on the basis of key words that are characteristic for the search query. Depending on the number of results or hits with which the presence of keywords in the short data records 10 is determined, the domain 4 corresponding to the appertaining short data record 10 is recognized as being relevant for the search query and the user is informed of the appropriate domain address on a list of results.

[0023] In order to generate the short data records 10 that are characteristic of the domains 4 and that in their totality form the index 8, the searching system 1 comprises a browsing module

16, also known as a "crawler". At regular, preferably cyclic time intervals, this browsing module 16 establishes contact with the appertaining domains 4 and browses them for their information contents. In this context, it can be particularly provided that the text information stored on a given domain 4 is acquired and appropriately compressed. The type and scope of the analysis of the content of a given domain 4 by the browsing module 16 are determined by the allocation of specific system resources of the browsing module 16 for the domain 4 in question. Here, the time span envisaged for the search, the computer capacity employed and/or the allocated memory capacity can all be specified as system resources as a function of the domain 4 in question. In particular, here it can also be specified whether that particular domain 4 is to be at all addressed by the browsing module 16 or whether it should be ignored altogether. Based on the information base acquired for that particular domain 4 during the browsing, the browsing module 16 generates the associated short data record 10 in the form of a summary and stores it as a component of the index 8 in the memory module 6.

[0024] The allocation of the system resources for browsing the domain 4 in question can be done, for example, as a function of domain-specific relevancy characteristic values. Here, so-called static relevancy characteristic values can also be provided which, on the basis of specifiable criteria such as, for instance, the degree of networking of the domain 4 with other domains 4, characterize the degree of acceptance of said domain 4 on the part of the users. Based on these relevancy characteristic values, it can then be determined whether a domain 4 is to considered at all during the browsing and, if so, how thoroughly this particular domain 4 should be browsed when the associated short data records 10 are being generated.

[0025] In addition, the searching system 1 is likewise configured to also take into consideration empirical values and knowledge from the preceding search queries when the short data records 10 are generated, thus also allowing the current user interests as reflected therein to be integrated to a large extent into the generation or cyclic updating of the index 8. For this purpose, an additional memory module 18 into which incoming search queries can be stored for further evaluation – along the lines of a logbook – is associated with the memory module 6. The contents of the memory module 18 are made available to an analysis module 20 that evaluates the incoming search queries and then, on the basis of the knowledge thus acquired, undertakes a

new distribution of the system resources to the domains 4 that are to be taken into consideration during the next browsing cycle. The appertaining allocation of the system resources is subsequently transmitted by the analysis module 20, as indicated by the arrow 22, to the browsing module 16.

[0026] Therefore, when the system resources are being allocated, the analysis module 20 considers empirical values from preceding search queries. This can be done, for instance, by determining the frequency of a search query or of a key word as an individual element of a search query, whereby frequently employed search queries or individual elements of search queries yield a conclusion about what is relatively popular among users at the present time. In a corresponding manner, it is assumed that the data records or domains 4 recognized as being relevant and found in relatively popular search queries reflect the current user interests to a fairly great extent. Therefore, in this embodiment, the analysis module 20 can allocate a correspondingly higher share of system resources to those domains 4 that were listed as the result of the relatively frequent employed search queries during the next browsing by the browsing module 16.

[0027] Moreover, the searching system 1 is additionally configured to also consider relatively complex structures in the profile of the search queries when the analysis module 20 allocates the system resources. In this process, especially correlations between individual elements of search queries are also taken into consideration when the empirical values are determined. For example, if it is determined that individual elements or search words appear in search queries particularly often combined with certain other individual elements or search words, it is concluded that a highly intrinsic correlation exists between these two search elements, so that, on the one hand, those domains 4 where complete or approximate combinations are found are recognized as being especially relevant while, on the other hand, when the relative frequency of individual search elements is assessed, the relative frequencies of the other search elements that correlate especially well with the former can also be taken into consideration.

[0028] For purposes of a statistical evaluation of the search queries, a correlation matrix is created in the analysis module 20, whereby the matrix elements of said correlation matrix are a quantitative measure of the correlation between two individual elements of search queries. In this context, especially the relative frequency with which each of the two individual elements of search queries are requested in combination with each other can be provided as a quantitative measure. This correlation matrix is subsequently diagonalized by a principal axis transformation, whereby the eigenvalues of the original correlation matrix are indicated on the main diagonal of the diagonalized matrix. The eigenvectors of the correlation matrix are also determined in this principal axis transformation.

[0029] Then the eigenvalues and eigenvectors of the correlation matrix can be employed for a further evaluation of the search queries. Those eigenvectors of the correlation matrix that exhibit a relatively large eigenvalue correspond to a mix of individual elements of search queries that shows up relatively often in typical search queries and thus said mix reflects the current user interest to a large extent. Therefore, in a subsequent step, those eigenvectors of the correlation matrix that are selected are those to which a relatively high eigenvalue is ascribed. The eigenvectors thus determined are obtained as a mix of search queries which are very highly probable to have appeared very recently in that particular combination.

[0030] By means of the thus selected "eigenqueries" associated with the relatively large eigenvalues of the correlation matrix, the analysis module 20 then accesses the index 8 in the form of a test query, thus determining for each "eigenquery" the data records or domains 4 that are recognized as being relevant for this eigenquery. Since the domains 4 determined in this manner correspond to a large extent to the current user interest, the system resources for these domains 4 are proportionally increased for the next browsing of the World Wide Web in comparison to the previous run. This can be done, for instance, by assigning a weighting factor when the system resources are allocated for the particular domain 4 according to the following equation

$$R_{VPA}(D_k) = (1 + \alpha * \lambda_\kappa^\beta)\alpha, \beta > 0$$

wherein the eigenvalue λ_κ of the appertaining eigenquery D_κ can be a domain 4 indicated as a hit for this eigenquery, and α can be a suitably selected constant > 0.

List of reference numerals

1 searching system
2 double-headed arrow
4 domain
6 memory module
8 index
10 short data records
12 arrow
14 input/output module
16 browsing module
18 memory module
20 analysis module
22 arrow